ANL-MCS-TM-369

Validation Study of CODES Dragonfly Network Model with Theta Cray XC System

Mathematics and Computer Science Division (MCS)

ANL-MCS-TM-369

Validation Study of CODES Dragonfly Network Model with Theta Cray XC System

prepared by
Misbah Mubarak
Robert B. Ross
Mathematics and Computer Science Division, Argonne National Laboratory

May 12th, 2017

Validation Study of CODES dragonfly network model with Theta Cray XC system[1]

Misbah Mubarak, Robert B. Ross and Christopher D. Carothers
MCS Division, Argonne National Laboratory, IL

This technical report describes the experiments performed to validate the MPI performance measurements reported by the CODES dragonfly network simulation with the Theta Cray XC system at the Argonne Leadership Computing Facility (ALCF).

1- Capturing performance data on Theta ALCF system:

Theta [1] is an Intel and Cray platform aimed to serve as the forerunner to the CORAL Aurora system [2]. It is a 9.65 Petaflops system with second generation Intel Xeon Phi processors and a high-radix dragonfly network topology [3]. Theta is equipped with 3,624 compute nodes with each node having 64 cores with 16GiB of MCDRAM, 192 GiB of DDR4 RAM and 128 GiB of SSD storage.

The dragonfly interconnect topology on Theta is similar to the one incorporated in Edison and Cori systems at NERSC [4,5]. Within each group, there are 64 routers laid out in the form of a 6x16 matrix. All of the 16 routers in a row are connected to each other using the green links (example connectivity is shown in Figure 1). All the routers in the same column are connected to each other using 3xblack links (example shown in Figure 1). Routers in a group are connected to routers in other group via blue links. On theta system, there are 24 global channels between each group where each router in the source group has two global channels with the same router in the destination group.

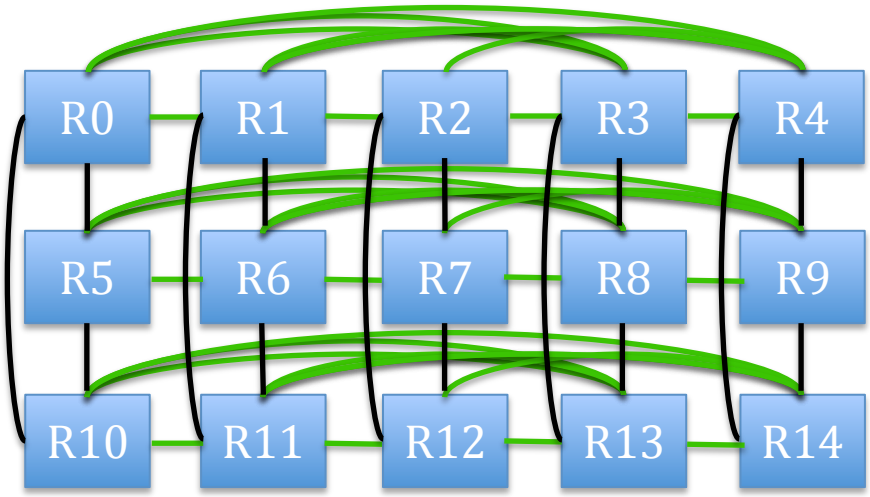


Figure 1: Example link arrangement within a dragonfly group for a 3x5 matrix.

[1] The publication of the technical report on the ANL website is currently in progress.

Communication traffic interference can be a reason to perturb the MPI performance measurements on dragonfly-based systems [6]. To eliminate the effect of communication traffic interference, the performance measurements were done after doing a full system reservation for 3,624 nodes on Theta. A mapping of one MPI rank per node was used in all the experiments since we were interested in the interconnect performance of the system.

2- Benchmarks used for performance measurements:

We used the following benchmarks for measuring the MPI performance on the Theta platform.

2.1 ***Ping Pong:*** The ping-pong benchmark sends MPI send/recv messages between two selected network nodes. In the ping-pong benchmark used on Theta, we recorded the performance of MPI blocking messages with size from 0 bytes to 65,536 bytes in increments of 1,024 bytes. Each message was sent 1500 times to avoid performance glitches.

To take into account the number of hops traversed by the messages, the ping-pong benchmark was configured to send messages between network nodes that are at varying distance to each other. Messages were exchanged between pairs of network nodes that were: (a) connected to the same router in a group (b) connected to different routers that are in the same row in a group (c) connected to routers in a group that do not have a direct connection in between (same group, one router hop in between the source and destination routers) (d) connected to routers belong to different groups.

2.2 ***Bisection Pairing:*** The bisection pairing benchmark is provided as part of the mpptest performance measurement suite [7]. As opposed to the ping –pong benchmark in which only two network nodes participate, the bisection pairing benchmark involves all network nodes that are part of the job by exchanging messages in pairs. Figure 2 demonstrates an example of how messages are exchanged between participating processes in a bisection pairing benchmark. The first and the last network nodes exchange MPI messages with each other while the network nodes in between exchange messages in pairs to generate network interference. Performance is reported for the first network node and the measurements were done for MPI messages having a size between 0 and 65,536 bytes. Mpptest sends each message 1400 times to avoid any performance glitches.

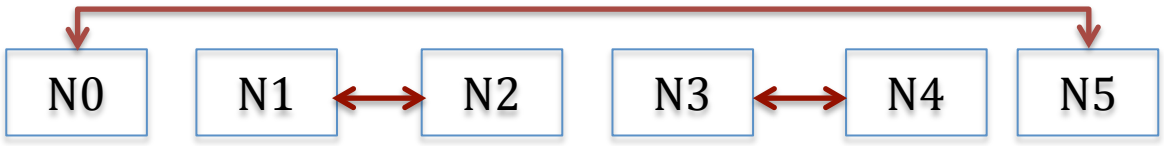


Figure 2: Bisection pairing benchmark for 6 network nodes

3- CODES Dragonfly Simulation Configuration:

The CODES simulation framework is able to ingest and replay the MPI traces generated by the DUMPI tracing library [8]. We recorded the DUMPI traces generated by the ping-pong and bisection pairing benchmarks on Theta. We then replayed the traces on the dragonfly network simulation that was setup using Theta's network configuration. The routing algorithm used was adaptive routing.

CODES provide MPI simulation layer that replays the MPI operations from the traces in the correct causality order. The MPI simulation layer provides the support for both eager and rendezvous protocols. The messages transferred via the eager protocol have an additional cost of buffer copying associated with them. Rendezvous protocol cuts down on the copying overhead by performing a handshake prior to the data transfer and making sure that sufficient buffer space is available at the receiving end.

For the performance measurements done on Theta, there was a performance variation at 8KiB message size at small scales and 2KiB message size at 1,024 nodes and beyond. Our conjecture was that prior to these points, the messages were being transferred using the MPI eager protocol and after this point, the messages were transferred using the rendezvous protocol. We therefore configured the simulation to switch to rendezvous protocol when doing message transfers of 8KiB at small scales and 2KiB at a scale of 1,024 and 2,048 nodes.

The simulation was configured with a MPI overhead of 1.25 microseconds, a NIC delay of 250 nanoseconds and a router delay of 100 ns. This overhead was derived by sending a zero byte message between nodes connected to the same router on Theta, which showed a latency of 3.1 microseconds (100 ns for traversing the router and 3 microseconds of base MPI and NIC overhead).

The overhead of copying message for MPI eager protocol is configured to 0.55 ns per byte. This cost per byte was derived by sending messages ranging from 0 to 1KiB using MPI eager protocol on Theta.

Throughout the simulation experiments, we map a single MPI rank to a network node similar to what we did for the performance measurements on Theta.

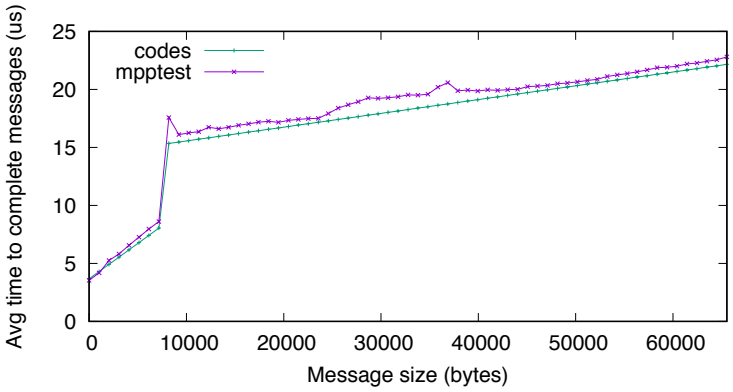
4- Validation Results:

We present the comparison of performance measurements done on Theta and the CODES dragonfly simulation configured using the parameters discussed in Section 3. The sharp performance variation is the point where protocols switch from eager to rendezvous.

4.1 Ping Pong benchmark

The comparisons of MPI performance measurements on Theta and the CODES dragonfly simulation using the ping-pong benchmark are shown in Figure 3 (a) –

(d). In all four cases, for majority of the message sizes the performance of CODES dragonfly model matches the measurements done on Theta. In some cases, there is less than 7% performance variation between the message latencies reported by CODES dragonfly simulation and Theta. We anticipate that these small variations are due to intra-node interference effects on Theta.



(a) Figure 3(a) Ping-pong message exchange between pair of nodes connected to same Router

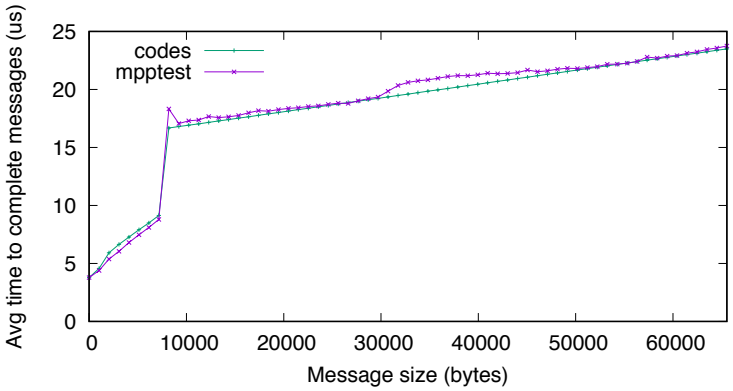


Figure 3 (b) Ping-pong message exchange between pair of nodes connected to different routers that are directly connected in the same group.

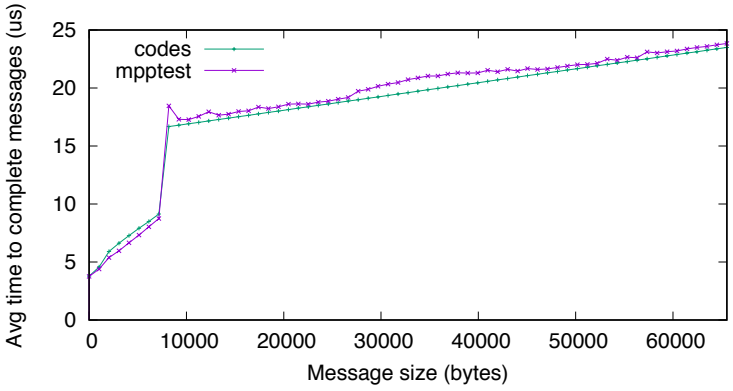


Figure 3 (c) Ping-pong message exchange between pair of nodes connected to different routers in a group that have no direct connection in between.

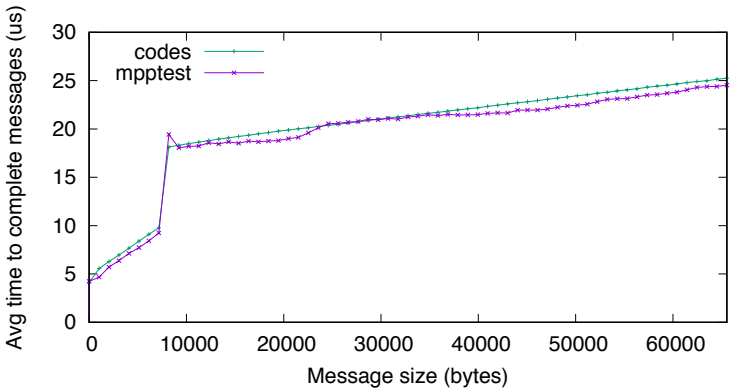


Figure 3 (d) Ping-pong message exchange between pair of nodes connected to different routers in different groups.

4.2 Bisection benchmark

The comparison of MPI performance measurements on Theta and the CODES dragonfly simulation using the bisection pairing benchmark is done using different network scale ranging from 4 nodes to 2,048 nodes. The performance results are shown in Figure 4 (a) – (f). In majority of the cases, the simulation performance matches the measurements recorded on Theta. In some cases, there is less than 8% performance variation between the message latencies reported by CODES dragonfly simulation and Theta. We anticipate that these small variations are due to minor differences in congestion sensing mechanism of adaptive routing and intra-node interference effects on Theta.

When scaling from 4 nodes to 2,048 nodes, there is a relatively small amount of difference in latency (on the order of 1-2 microseconds). There are multiple reasons that account for this small difference. First, both the simulation and validation tests are run on a quiet system with no background traffic interference. Second, the cost of traversing an intermediate router is much lower (100ns) as compared to other hardware/software overheads from MPI and the NIC (3 microseconds). Third, there are multiple routes involving different local and global channels from the source and destination and if adaptive routing does effective load balancing then the impact of congestion is minimized.

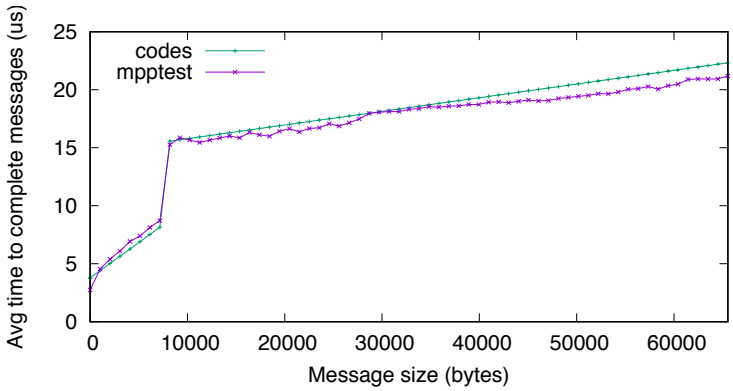


Figure 4 (a) Bisection pairing message exchange for 4 nodes.

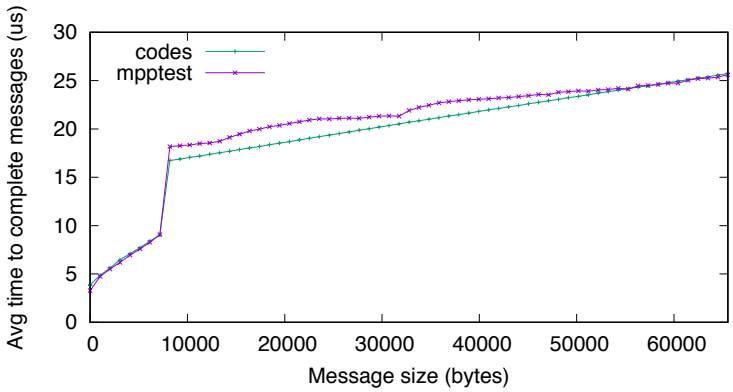


Figure 4 (b) Bisection pairing message exchange for 16 nodes.

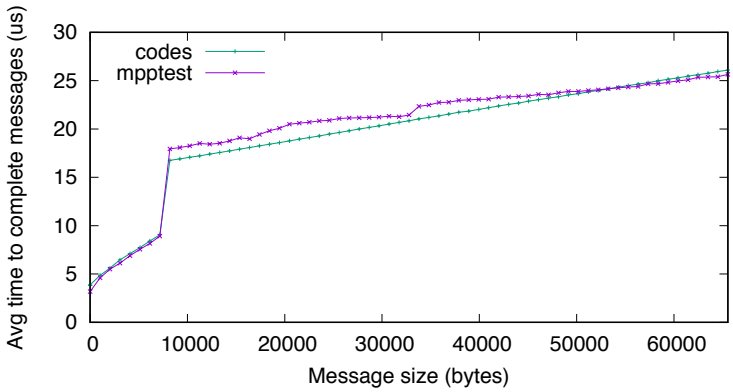


Figure 4 (c) Bisection pairing message exchange for 64 nodes.

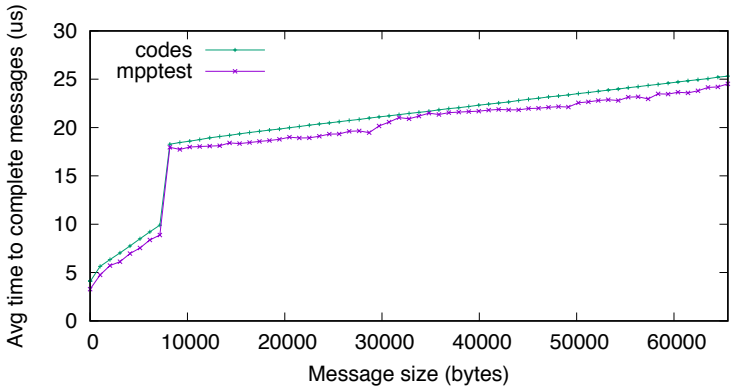


Figure 4 (d) Bisection pairing message exchange for 256 nodes.

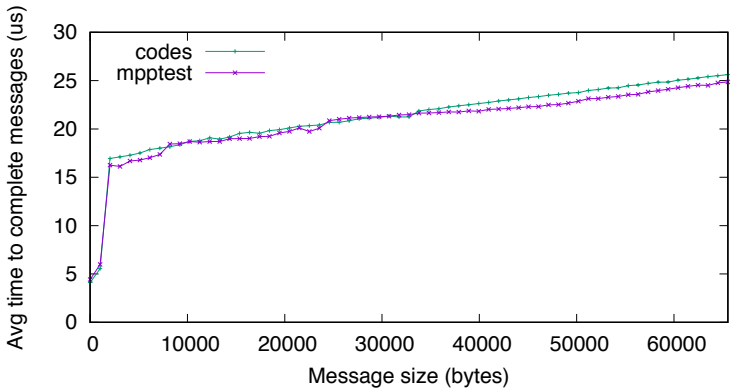


Figure 4 (e) Bisection pairing message exchange for 1,024 nodes.

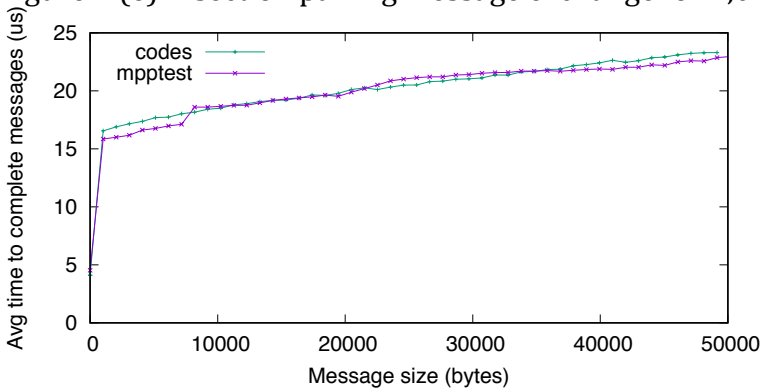


Figure 4 (f) Bisection pairing message exchange for 2,048 nodes (50KiB message sizes are used with 2,048-node simulation run).

References:

[1] Theta Argonne Leadership Computing Facility: https://www.alcf.anl.gov/theta

[2] Aurora Argonne Leadership Computing Facility: http://aurora.alcf.anl.gov

[3] G.Faanes,A.Bataineh,D.Roweth,T.Court,E.Froese,B.Alverson, T. Johnson, J. Kopnick, M. Higgins, and J. Reinhard, "Cray cascade: A scalable hpc system based on a dragonfly network," in *High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for*, Nov 2012, pp. 1–9.

[4] Edison Cray XC30 System at NERSC: http://www.nersc.gov/users/computational-systems/edison/

[5] Cori Cray XC40 System at NERSC: http://www.nersc.gov/users/computational-systems/cori/

[6] X. Yang, J. Jenkins, M. Mubarak, R. B. Ross, and Z. Lan. Watch out for the bully! job interference study on dragonfly network. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, ser. SC, volume 16, 2016.

[7] Mpptest performance measurement: http://www.mcs.anl.gov/research/projects/mpi/mpptest/

[8] DUMPI tracing tool: http://sst.sandia.gov/using_dumpi.html

Mathematics and Computer Science Division
Argonne National Laboratory
9700 South Cass Avenue, Bldg. 240
Argonne, IL 60439

www.anl.gov

Argonne National Laboratory is a U.S. Department of Energy
laboratory managed by UChicago Argonne, LLC